**Probabilistic Sufficient Explanations**

**Motivation**

We want to generate local explanations for a given classifier.

- **Logical Reasoning**: Aims for 100% guarantee. Too strict, can result in complex explanations.
- **Model Agnostic**: Hard to capture the dependencies between features. Generally, ignore feature distribution (can be fooled). For example, LIME and SHAP are in this category.

**Proposed Solution**: Choose a subset of given features and treat the rest as missing. Also want to provide some probabilistic guarantees about the outcome of the classifier while prioritizing “simpler” subsets.

**Probabilistic Sufficiency**

Two intuitive metrics which can be used to evaluate explanation quality.

- **Same Decision Probability** (SDP): probability the classifier will make the same decision when observing the rest of the features.

\[
SDP_{C,x}(z) = \frac{\mathbb{E}_{m \sim P_t(M|x)} \left[ C(z|m) = C(x) \right]}{C(x)}, \quad \Rightarrow \text{intractable to compute} \ [2, 1]
\]

- **Expected Prediction** (EP): how “confident” the classifier is on its decision.

\[
EP(z) = \frac{\mathbb{E}_{m \sim P_t(M|x)} f(z|m)}{P_t(M|x)}, \quad \Rightarrow \text{for some distribution, classifier pairs} \ [3, 4, 5]
\]

Connection between SDP and EP:

\[
SDP_{C,x}(z) > \frac{EP(z) - T}{U(z) - T}.
\]

**Probabilistic Sufficient Explanations**

**Idea**: We want to provide good probabilistic guarantees while choosing a small subset of features.

We define the sufficient explanations to be

\[
SE_k(x) = \arg \max_{z \subseteq x} EP(z) \quad \text{s.t.} \ |z| \leq k
\]

Out of these, we want the most likely ones:

\[
MLSE_k(x) = \arg \max_{z \in SE_k(x)} P_t(z)
\]

**Finding Sufficient Explanations**

We use beam search algorithm guided by expected prediction to greedily find the subset of features that give us best guarantee. The iterative nature of beam search allows us to save explanations of different sizes.

**Experiments**

MNIST 3 vs 5 binary classification using decision forest classifier and probabilistic circuit for feature distribution

![Correctly classified examples](image1)

![Misclassified examples](image2)

From left to right: original image, anchors, ours (same size), ours (size 30)

- Chosen pixels mostly in upper part of image - where 3’s and 5’s differ
- White pixels show outline of predicted number; black pixels where the other number may be present

<table>
<thead>
<tr>
<th>Method</th>
<th>EP(0)</th>
<th>SDP(3,5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchors</td>
<td>0.75 ± 0.37</td>
<td>0.66 ± 0.08</td>
</tr>
<tr>
<td>MLSE_5</td>
<td>1.57 ± 0.29</td>
<td>0.86 ± 0.05</td>
</tr>
<tr>
<td>MLSE_30</td>
<td>3.75 ± 0.13</td>
<td>1.00 ± 0.00</td>
</tr>
</tbody>
</table>

\[\Rightarrow \text{Our explanations have high expected predictions and high SDP}\]

\(\Rightarrow\) squeezing out small gains in expected prediction results in much less likely (more complex) explanations

**References**


